J. Mol. Biol. (1997) 271, 511–523

JMB

Correlated Mutations Contain Information About Protein–protein Interaction

Florencio Pazos[1], Manuela Helmer-Citterich[2], Gabriele Ausiello[2] and Alfonso Valencia[1]*

[1]*Protein Design Group CNB-CSIC, Campus U. Autónoma, Cantoblanco Madrid 28049, Spain*

[2]*Dipartimento di Biologia U. di Roma "Tor Vergata" Rome, Italy*

**Corresponding author*

Many proteins have evolved to form specific molecular complexes and the specificity of this interaction is essential for their function. The network of the necessary inter-residue contacts must consequently constrain the protein sequences to some extent. In other words, the sequence of an interacting protein must reflect the consequence of this process of adaptation. It is reasonable to assume that the sequence changes accumulated during the evolution of one of the interacting proteins must be compensated by changes in the other.

Here we apply a method for detecting correlated changes in multiple sequence alignments to a set of interacting protein domains and show that positions where changes occur in a correlated fashion in the two interacting molecules tend to be close to the protein–protein interfaces. This leads to the possibility of developing a method for predicting contacting pairs of residues from the sequence alone. Such a method would not need the knowledge of the structure of the interacting proteins, and hence would be both radically different and more widely applicable than traditional docking methods.

We indeed demonstrate here that the information about correlated sequence changes is sufficient to single out the right inter-domain docking solution amongst many wrong alternatives of two-domain proteins. The same approach is also used here in one case (haemoglobin) where we attempt to predict the interface of two different proteins rather than two protein domains. Finally, we report here a prediction about the inter-domain contact regions of the heat- shock protein Hsc70 based only on sequence information.

Keywords: correlated mutations; protein contacts; docking; co-adaptation; Hsc70

Introduction

The protein–protein interaction problem

Molecular recognition is a key process in biological systems. The order and control of protein–protein interactions in signalling pathways and metabolic networks are important aspects of molecular biology and biochemistry. DNA replication and transcription, RNA splicing, protein sorting, cell adhesion, signalling cascades and metabolic cycles are just some examples of the many complex processes dominated by protein–protein recognition.

The unravelling of this complex process requires two major steps. First, it is necessary to find the interacting proteins in the cell soup (e.g. the long quest for the downstream effectors in the ras-p21 signalling cascade); and second, to describe at the molecular level how the interaction takes place, e.g. how and where ras-p21 interacts with the raf-kinase and how conformational changes related with GTP hydrolysis control the interaction between the proteins.

The first issue of searching for the interacting components in a functional complex is a daily problem for experimental biology but has remained so far untouched from the theoretical point of view. The problem of describing and predicting the mol-

E-mail: valencia@cmb.uam.es

Abbreviations used: RMS, root-mean-square deviation; *Xd*, harmonic difference between binned populations; Nt, N-terminal; Ct, C-terminal.

0022–2836/97/340511–13 $25.00/0/mb971198

ecular complexes in detail, also known as the "docking problem" has instead attracted a great deal of attention and has led to the development of several different theoretical methods.

Current physical approaches to the docking problem

Docking has attracted much attention (for recent reviews, see Lengauer & Rarey, 1996; Strynadka *et al.*, 1996). Undoubtedly, progress has been made, and some methods are ready for challenges such as the prediction of the interaction between lactamase and one of its inhibitors. It was accomplished quite successfully by six different groups (Strynadka *et al.*, 1996), or the more recent CASP-2 meeting (WWW: http://iris4.carb.nist.gov/casp2). All current docking methods require the three-dimensional structures of the interacting proteins to be known. In all methods, the interacting surfaces are described by different physical properties (Connolly surfaces, grids, protein slices, property vectors, etc.) to allow the identification of geometrically complementary regions between the two proteins (Cherfils *et al.*, 1991; Fischer *et al.*, 1995; Helmer-Citterich & Tramontano, 1994; Jackson & Sternberg, 1995; Jiang & Kim, 1991; Shoichet & Kuntz, 1991; Stoddard & Koshland, 1992; Walls & Sternberg, 1992; Katchalski-Katzir *et al.*, 1992).

Most algorithms treat proteins as rigid bodies and in only a few cases is protein flexibility taken into account (Totrov & Abagyan, 1994; O'Donoghue, S. & Nilges, M., personal communication). Flexibility is an inherent difficulty in the docking problem, since most inter-protein complexes undergo induced-fit movements upon binding, and hence a rigid-body description of the individual components may not be accurate enough to predict the structure of the final complex.

Docking methods and the characteristics of protein interfaces

It is generally accepted that the physical principles underlying protein folding and protein–protein association are similar. This belief is supported by detailed studies of similarities in the packing of protein interfaces and protein interiors (Walls & Sternberg, 1992), and similarities in the overall resemblance of the hydrophobicity patterns (Young *et al.*, 1994). However, our understanding of the peculiar characteristics of protein-protein interaction is still very limited. Earlier attempts to study complementary surfaces between proteins (Argos, 1988; Janin & Chothia, 1990; Janin *et al.*, 1988) were hampered by the lack of experimental data. More recent studies (Jones & Thornton, 1996; Tsai *et al.*, 1996) are for the first time providing tools for a systematic approach to the characterisation of protein–protein interfaces by rigorous scanning of data bases of protein complexes.

The study of the evolution of oligomerisation has also become an important issue and the first ideas about the origin of the adaptation in protein complexes from the initial components are emerging (Fletterick & Bazan, 1995; Bennett *et al.*, 1995).

A new approach to predict protein–protein contact regions based on sequence information

We propose here a new and completely different approach to the study and prediction of protein–protein interaction. Instead of considering the structural nature of the interactions, we try to detect the sequence traces that evolution may have left on the interacting sequences during the process of preserving the protein–protein interaction sites. Therefore, our approach is not restricted to the cases in which the structures of the proteins to be docked are known and is applicable to any family of interacting proteins for which a large enough sequence family is available.

Sequence information and the process of protein–protein co-adaptation

There is a common agreement among researchers that interacting proteins undergo a process of co-evolution. "Over time, amino acid substitution may stabilise an interface that does not exist in the closed monomer ... stabilising mutations in these interfaces would be favoured in natural selection" (Bennett *et al.*, 1995); however, no explicit strategy has been proposed for detecting the traces of this process from protein sequences. We propose that it is possible to detect this signal by studying compensatory mutations. In order to do so, we have appropriately modified our previously published method for the calculation of correlated mutations in multiple-sequence alignments (Göbel *et al.*, 1994; Pazos *et al.*, 1997).

Defining correlated mutations

Several groups have studied correlated mutations: technical differences between different approaches have led to conflicting conclusions about the nature and intensity of this phenomena (Altschuh *et al.*, 1987, 1988; Göbel *et al.*, 1994; Neher, 1994; Shindyalov *et al.*, 1994; Taylor & Hatrick, 1994). Thus, it is important to define exactly our notion of correlated mutations. In this and previous work, we have used the term correlated mutations to indicate a tendency of positions in proteins to mutate co-ordinately. We measure this tendency by analysing the correlation between changes in pairs of positions in multiple sequence alignments, with an unambiguous definition of correlation (see Methods).

Biological meaning f compensatory mutati ns

There is clear experimental support for the role of compensating mutations in protein stability

(Serrano *et al.*, 1990) and functi n (Gregoret & Sauer, 1993). Vernet *et al.* (1992) directly tested the influence in protein stability of some pairs of correlated mutations. We believe that the signal detected by our method corresponds mainly to networks of positions that have undergone compensating mutations during evolution. If interactions between proteins are of the same physical nature as intra-protein interactions, then their consequences at the sequence level are most likely also similar. Therefore, we apply our method, which we have previously used to predict contacts in globular proteins (Göbel *et al.*, 1994; Pazos *et al.*, 1997), to the problem of predicting interactions between proteins. As we will show here, the signal at the sequence level for inter-protein contacts turns out to be even more specific than that for intra-protein contacts, possibly because it is subject to a stronger selective pressure.

Testing the method

The purpose of this work is to test the feasibility of a sequence-based approach to the prediction of interacting regions in protein complexes. To do so, we first show that correlated pairs between two different proteins are significantly closer to each other than other pairs of positions in the same proteins, and second that they can be used to discriminate the correct docking solutions among many alternative wrong ones in proteins of known structure. We then carry out a *bona fide* prediction of the yet unknown interaction site between the two domains of the Hsc70 heat shock protein

Results

The results are presented in the following order: first we demonstrate that correlated mutations do contain information about inter-domain contacts. We tested our method mainly for inter-domain interactions to take advantage of the larger set of examples of proteins of known structure for which many homologous sequences are available. As described in detail later, we demonstrate that, on average, pairs of residues detected as "correlated" by our method are closer to each other than the average pairs of residues in the same protein.

Next we show how the correlated mutation analysis can be used to identify docking solutions very close to the native solution from many wrong solutions. The aim of this experiment is to empirically evaluate how much information about inter-domain contacts is contained in correlated mutations. It is important to remember that we are not attempting to replace existing docking methods based on structural information; we only want to establish that correlated mutations are good indicat rs of contacting residues. Our results should not be compared to any current docking method. The docking algorithm here is used only as a rapid tool to generate many alternative "reasonable" solutions. In a first set of experiments, we generated thousands of alternative solutions that fully cover the space of possible solutions without any attempt to increase the number of solutions close to the real docking position. The second set of experiments corresponds to a more demanding test, since in this case the set of solutions among which we want to discriminate consists of hundreds of physically realistic solutions, corresponding to the best scoring complementary surfaces calculated with a standard docking program.

The third section contains the results obtained for the complex between α and β haemoglobin, a test selected to prove that our method has similar performances when applied to inter-protein as well as inter-domain contacts. A more complete test of protein–protein complexes is prevented by the very limited availability of data on protein–protein complexes where both the structure and a sufficient number of aligned sequences for the same species (see later) are known.

Finally, a *bona fide* prediction is reported. Sequence information is used to predict the contacting residues between the two domains of the heat-shock protein Hsc70. This prediction is used to illustrate the novel feature of our approach: that prediction can be made in the absence of structural information. The example is also biologically relevant: the function of Hsc70 is based on the interaction between its two domains. The N-terminal (Nt) domain contains the ATP-binding site, while the C-terminal (Ct) domain is mainly responsible for peptide binding (Chappel *et al.*, 1987; Gragerov *et al.*, 1994; Montgomery *et al.*, 1993). The interaction between the two domains generates the biological functions of peptide binding and release (McCarty *et al.*, 1995). The structure of both isolated domains (Nt domain of hsc70; Flaherty *et al.*, 1990; Ct domain of its related protein DnaK, Zhu *et al.*, 1996) has been solved, although the Ct domain structure is not yet publicly available. This is an appropriate moment for a *bona fide* prediction, since the structure of the complex has not yet been solved and "classical" docking methods cannot be used until both structures are available.

Prediction of domain–domain contacts for different protein families

We have previously shown that in single-domain proteins correlated residues tend to be closer than other residues (Göbel *et al.*, 1994). This general result is illustrated in Figure 1(a) for papain (9pap, Kamphuis *et al.*, 1984). In the Figure we compare the distribution of the distances between pairs of correlated residues with that of the distances between all pairs of residues in each f the domains of papain. The distribution shows a clear shift of the population of correlated positions toward closer distances. This example supports our earlier conclusion that, in globular proteins, correlated

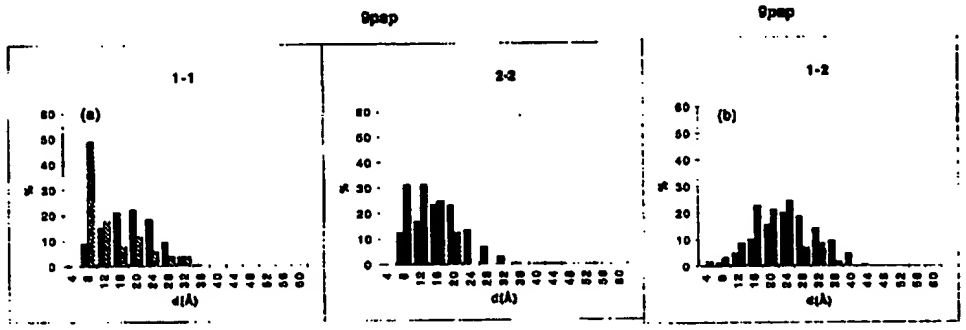

Figure 1. Bar diagrams comparing the proportions of pairs of residues at different distances. Distributions are represented for all residues (filled bars) and for correlated pairs of residues (hatched bars) in papain (9pap). (a) Distances between pairs in the two independent domains, and (b) distances between the two domains. Correlated positions are shifted toward smaller distances.

positions are statistically closer than non-correlated positions.

The spatial proximity of correlated positions inside globular proteins can be extended to the proximity of correlated positions belonging to two different domains. The actual values of inter-domain distances for all residues and for correlated positions belonging to two different domains for papain are compared in Figure 1(b). Once again a clear shift of the correlated pairs toward smaller distances is observed. This shift indicates that correlated positions have a tendency to be closer to

Table 1. The 21 monomeric protein families with two domains

PDB code[a]	Size[b]	Domain Size[c]	Domain definition[d]	HSSP NALIGN[e]	*Xd* 1/2 I-I[f]	*Xd* 1/2 II-II[g]	*Xd* 1/2 I-II[h]	Reference[i]
A. *Disjoin domains*								
4mt2	61	29/32	1–29/30–61	72	Nep	2,36	−2,22	(2)
3dfr	162	80/68	2–32, 112–160/38–105	18	11,68	7,92	−1,63	(3)
4tnc	162	88/72	3–90/91–162	139	4,39	10,16	0,45	(2)
3cln	148	74/66	5–78/82–147	167	2,57	8,61	1,63	(3)
1clm	148	85/59	4–88/89–147	175	1,58	10,51	2,19	(1)
B *Conjoin domains*								
1rnd	124	73/51	1–49,80–103/50–79, 104–124	62	7,57	3,56	−0,77	(1)
4tms	316	223/93	1–52,146–316/53–145	20	4,39	Ne	1,50	(1)
3pgk	416	199/216	1–188,405–415/189–404	42	5,07	1,95	2,36	(1)
C. *Interacting domains*								
2gcr	173	82/92	1–82/83–174	45	5,22	16,71	0,81	(3)
1alc	123	67/55	38–104/1–37,105–122	67	2,40	4,26	2,53	(2)
3blm	257	156/86	1–67,168–256/69–154	33	2,94	9,66	3,72	(3)
2pf2	156	62/83	1–62/63–145	20	2,39	5,91	4,05	(1)
2bbm	148	74/66	5–78/82–147	177	4,42	9,34	4,31	(3)
1ppl	323	212/111	1–192,304–323/193–303	30	4,39	−1,14	4,53	(1)
3est	240	126/103	16–29,122–233/30–120,234–245	211	18,35	18,87	5,27	(1)
3adk	195	144/50	1–37,88–194/38–87	34	13,70	−6,87	6,72	(1)
3rp2	224	109/108	16–21,128–230/28–122,231–243	181	12,30	13,37	7,10	(3)
9pap	212	112/100	1–16,113–208/17–112,209–212	68	18,32	7,98	7,12	(1)
2c2c	112	50/62	1–33,96–112/34–95	117	8,60	8,44	8,85	(2)
3trx	105	72/35	1–72/74–108	45	3,91	8,46	9,53	(3)
1sgt	223	139/80	16–28,69–80,121–234/29–68,81–120	158	19,37	15,00	12,70	(1)

Nep, Not enough pairs for *Xd* calculation.

[a] PDB and chain identifiers.

[b] Chain length (in amino acid residues).

[c] Length of each domain (in amino acid residues).

[d] Domain definitions (DI/DII).

[e] Number of sequences in the multiple sequence alignment. From the HSSP Data Base (Sander & Schneider, 1993).

[f] *Xd*, the harmonic weighted difference between the binned distributions of distances between all residues and correlated pairs of positions in the first domain (see Methods).

[g] As [f] but for residue pairs in the second domain.

[h] As [f] but for pairs of residues belonging to different domains.

[i] The domain definition was taken from: (1) Holm & Sander (1994); (2) Siddiqui & Barton (1995); (3) Sowdhamini & Blundell (1994); (4) Swindells (1994).

the inter-domain interface. In order to quantify the difference between the population of distances between correlated pairs and that of all other pairs, we define the parameter *Xd*, as the harmonic difference between the two binned populations (see Methods).

Table 1 shows the results for a set of 21 proteins with two domains. In almost all cases, the population of correlated pairs inside the individual domains (domain I or domain II) is shifted toward smaller distances as indicated by large positive *Xd* values. There are four exceptions with negative or very small *Xd* values. We suspect that this may be due to imprecise definitions of domain boundaries leading to non-perfectly globular proteins, with distorted residue distance distributions.

Table 1 shows that correlated positions between the two protein domains are, on average, closer than non-correlated positions, with positive *Xd* values for 17 out of the 21 cases. In this Table, proteins are divided into three categories according to their degree of interaction, from weakly to strongly interacting domains (disjoin, conjoin and interacting, following Sowdhamini & Blundell (1994)). Correlated positions are closer only when there is a real association between domains, as demonstrated by higher *Xd* values for interacting domains than for any of the other two categories.

One particular example may illustrate the differences found in the analysis of domains with different degree of the interaction. In two crystal forms of calmodulin (PDB code 3cln (Babu *et al.*, 1988) and 1clm (Rao *et al.*, 1993)) the two protein domains do not interact, the closest residues belonging to different domains are more than 24 Å apart. When using this structure to calculate the distance distributions, correlated mutations cannot be much closer than the rest of the residues, and in fact the *Xd* value is only 2.19 (Table 1 and Figure 2(a)). These crystal structures represent only one of two alternative forms of calmodulin. A different form is observed (2bbm, Ikura *et al.*, 1992) where the two domains embrace a bound peptide substrate. When this "closed" form of calmodulin is used to calculate distances, correlated residues are closer than non-correlated ones with an *Xd* value of 4.53 (Table 1 and Figure 2(b)). This value is in good agreement with that measured for weakly-interacting domains of other proteins. In other words, in calmodulin, correlated mutations contain information about inter-domain contacts, and these are related to the interactions observed in the closed structure.

The observation that correlated positions tend to be closer to domain interfaces only in truly interacting domains could be interpreted as an indication of the physical nature of correlated positions as compensatory mutations, since compensation can occur only when physical contact between interacting residues has led them to co-evolve.

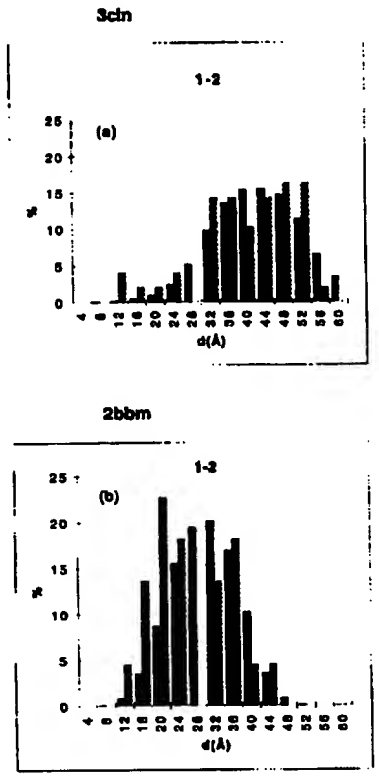

Figure 2. Bar diagrams comparing the proportions of pairs of residues at different distances in two conformational states of calmodulin. The distribution of distances between all pairs of residues (filled bars) and correlated pairs (hatched bars) are compared for the open ((a) 3cln) and closed ((b) 2bbm) forms of calmodulin. There is no interaction between the two domains in the open form (disjoin according to Sowdhamini & Blundell, 1994) and a moderate interaction in the closed form (classified as interacting). The population of correlated mutations is shifted more obviously toward shorter distances when the closed form is used to calculate the distances. We obtained *Xd* values of 4.31 and 2.19 for the closed and open form, respectively. Notice in the particular example of (b) there are not correlated pairs at distances between 24 and 28 Å.

The information about correlated positions may be sufficient for selecting the correct inter-domain docking solutions among many alternative possibilities

We generated a large number (7440) of random solutions for the docking of two interacting protein domains (Figure 3) in order to cover as thoroughly and evenly as possible the protein surface with physically realistic solutions, and tried to singl out th correct orientation using correlated mutations. The right docking solutions are clearly distinguished from the random alternatives (Figure 3).

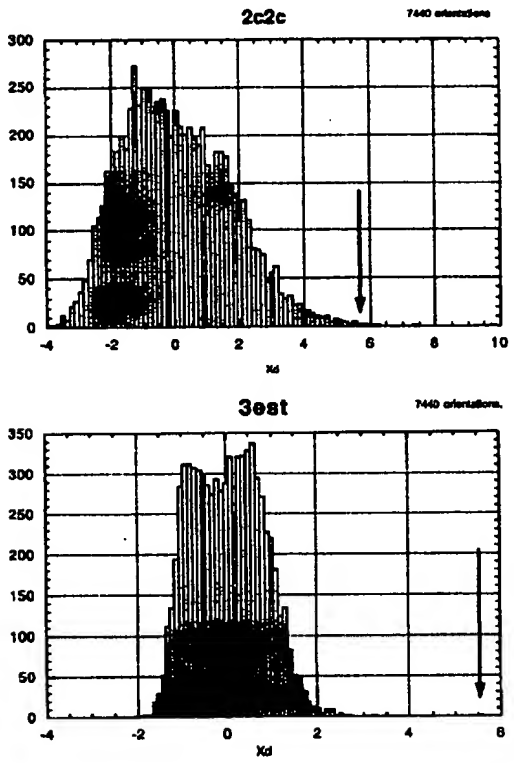

Figure 3. Distribution of the *Xd* values obtained for 7440 docking solutions covering the full docking space for the two domains of cytochrome *c* and trypsin. The *x*-axis, harmonical difference (*Xd*) between the distributions of relative distances between all pairs of residues and correlated pairs of positions (see Methods). The *y*-axis is the number of docking solutions corresponding to each *Xd* value. (a) 2c2c, cytochrome, (b) 3est, trypsin. The different solutions cover all the possible range of interactions between the two domains and they have been selected to be physically realistic in surface complementarity.

Because of the nature of our calculation, we are able to detect the solutions that best fit the observed pattern of correlation. This implies that, even if the right solution was not present in our set, the solutions closer to it would be identifiable. In fact, correlated positions tend to be found closer in space also in non-optimal solutions whose docking orientation is similar to the correct one.

To illustrate the first point, we compared the *Xd* values for the best solutions (lower RMS from the correct orientation) with the rest of the distribution. Even if these solutions are relatively far apart from the real solution (average RMS values of 10.02 and 14.07 Å) they have high *Xd* values (averages of 3.06 and 2.11). With these *Xd* values, they can be clearly distinguished from the bulk of the random solutions. Conversely, the ten solutions having a higher *Xd* value hav average RMS values of 27.24 and 20.97, clearly smaller than the average values for all other random s lutions (34.88 and 45.44 Å, respectively).

We also designed a second more demanding experiment. Instead of random docking solutions we chose the best solutions generated with a standard docking program based on surface complementarity (ESCHER, see Methods). Each one of these solutions is an alternative to the real solution and corresponds to local optima of surface complementarity (see Methods).

The results for all the proteins with strong inter-domain interaction in Table 1 (labelled as interacting domains) are given in Figure 4. First, there is a general correlation between RMS and discrimination by correlated mutations (*Xd*). With two exceptions (2gcr and 3adk), correct docking solutions with low RMS values have larger *Xd* values than other docking solutions with larger RMS values.

Results are quantified in Table 2. Correct docking solutions are always among the best 15% of all solutions, with the two exceptions mentioned above. We define as correct those solutions with an RMS lower than 5 Å from the experimental structure, since they represent only small differences in rotation of the two domains; this can be considered as the limits of resolution of the sequence-based method. Using this more relaxed criteria, then at least one right docking solution is found among the 8% best *Xd* values in 11 of the 14 cases.

For three examples, sequence information does not seem to be sufficient to discriminate among alternative docking solutions; the two mentioned before (2gcr and 3adk) and 2pf2 (where the docking method did not find any valid alternative solution close to the experimental structure). After visual inspection our interpretation is that highly symmetrical or elongated complexes are difficult for our method. It is imaginable that in a very elongated thin protein alternative docking in two opposed faces will give very high RMS values, while the distance between correlated pairs may be very similar in both of them.

Figure 5(a) shows some of the solutions for papain (9pap). The solution at 5 Å RMS is very close to the real solution for any biological application. A solution of high RMS shows how correlated pairs are clearly sitting in different faces of the protein. The structure of a wrong docking solution that scores better than the real solution shows how correlated residues are marginally closer than in the right orientation due to the particular shape of the region in which most of the correlated residues are found.

A further step: correlation between two different proteins

The same idea, applied abov to protein domains, can be used to predict the interaction between different proteins. There is a practical difficulty in testing this idea: currently there are

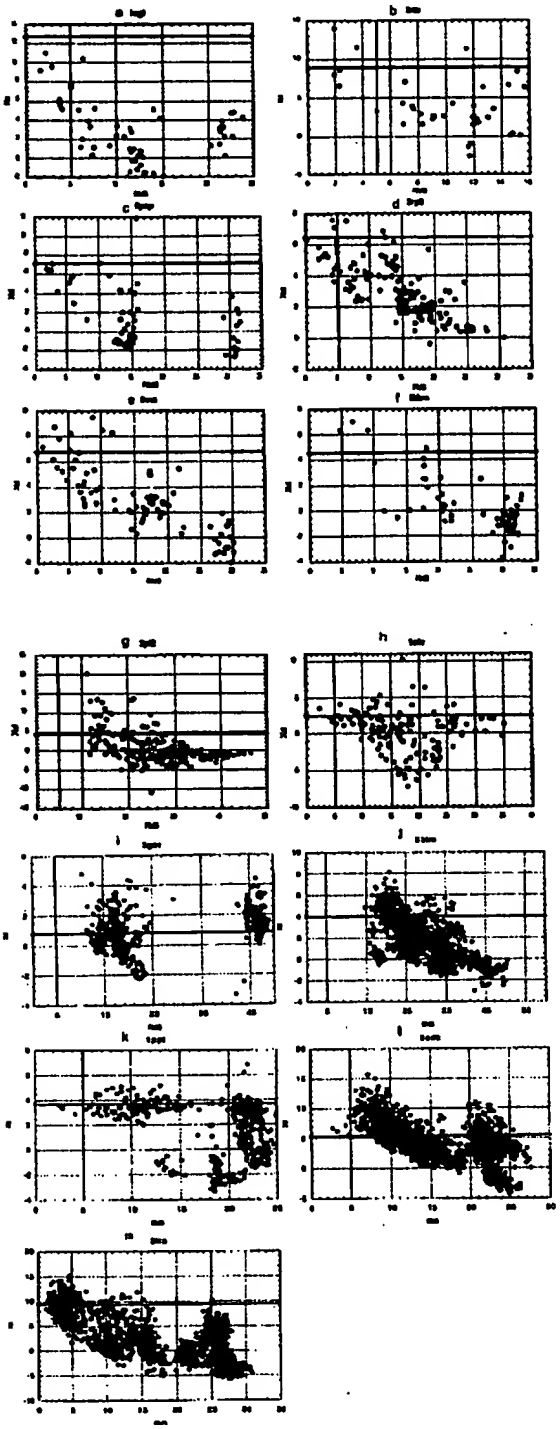

Figure 4. Scatter plot of the values of *Xd* against RMS. The *y*-axis, harmonic difference (*Xd*) between the distributions of relative distances between all residues and correlated pairs of positions (see Methods). The *x*-axis is the RMS between the real structure and the different alternative docking positions. (a) 1sgt, trypsin; (b) 2c2c, cytochrome; (c) 9pap, papain; (d) 3rp2, rat mast cell protease; (e) 3est, elastase; (f) 2bbm, calmodulin bound to substrate; (g) 2pf2, prothrombin; (h) 1alc, α-lactoalbumin; (i) 2gcr, γ-crystallin; (j) 3blm, β-lactamase; (k) 1ppl, penicillopepsin; (l) 3adk, adenylate kinase; (m) 3trx, thioredoxin. The different examples cover all the range of interacting domains. The right solution is always selected among the 8% best docking solutions, except for 2gcr, 3blm, 2pf2 and 3adk (see Table 2). Solutions with RMS smaller than 5 Å can be considered as valid solutions for the level of resolutions expected from a method based only on sequences (vertical thick line on the plots).

Table 2. Percentage of docking solutions that score better than the correct one for 11 proteins with interacting domains

PDBId	%[a]	%(5 Å)[b]
2gcr	53.49	53.49
1alc	15.07	7.95
3blm	13.90	13.90
2pf2	13.76	13.76
2bbm	6.25	3.08
1ppl	2.90	2.90
3est	9.76	1.22
3adk	28.73	28.48
3rp2	2.53	0.00
9pap	3.08	3.08
2c2c	7.69	0.00
3trx	7.75	2.46
1sgt	0.00	0.00

[a] Percentage of solutions scored better than the real one.
[b] The same considering all the solutions in the range 0 to 5 Å RMS as correct.

few cases where the three-dimensional structure of the protein complex and many corresponding sequences from different species are available.

We have chosen to test the concept using haemoglobin, since the wealth of sequences for this protein made it possible to select an appropriate subset of sequences. We selected the α1-β2 monomers of the tetramer because they contain the functional interface that undergoes structural changes between the oxy and deoxy forms (Jayaraman *et al.*, 1995). We generated multiple docking solutions for the α1-β2 dimer and used, as in previous cases, the information about correlated mutations to discriminate among them. As can be seen in Figure 5(b), there is a clear difference between good solutions and wrong ones. *Xd* values can be used to discriminate between them: the right solution always scores among the first 6% and the *Xd* value is 3.36, in the range of the examples of interacting domains shown in Table 1. This result is satisfactory, especially in view of the fact that this is a difficult case with a very small interacting surface (Lesk & Chothia, 1980; Perutz, 1978).

Prediction of contact between domains in the absence of three-dimensional structures

To illustrate the predictive value of our method we present a prediction of the domain interaction

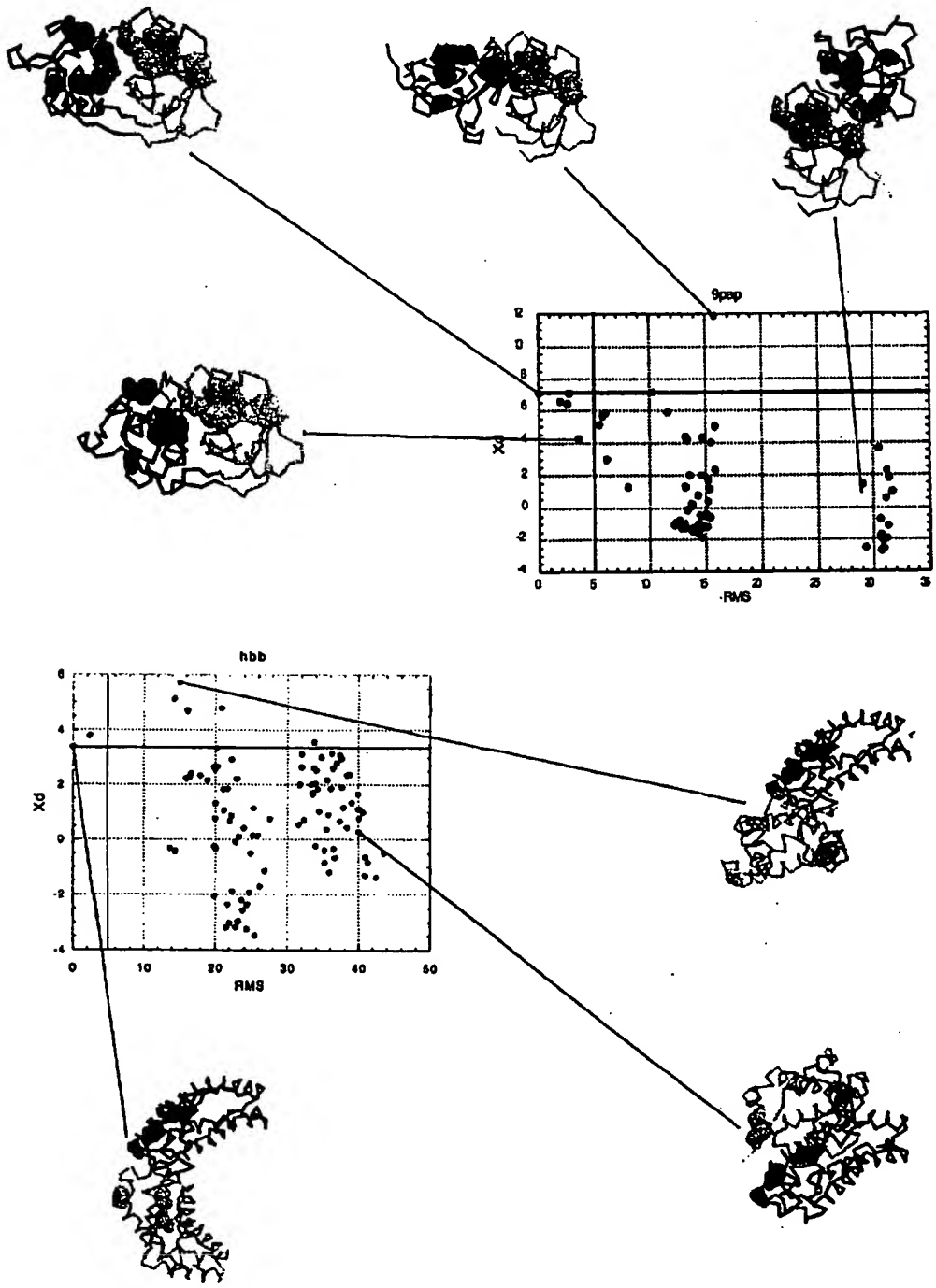

Figure 5. Scatter plot of the values of *Xd* against RMS for different docking solutions of (a) 9pap and (b) α1-β2 of haemoglobin (1hbb, Kavanaugh *et al.*, 1992). The *x*-axis, RMS between the real structure and the different alternative docking positions. The *y*-axis, harmonical difference (*Xd*) between the distributions of relative distances of all residues and correlated pairs. The right docking solution is among the 5% best scored ones. In the case of haemoglobin it is difficult to generate alternative docking solutions close to the real one, since the surface of interaction of the monomers is sharp and small. A ribbon representation of some docking solutions including the real one is shown. The residues participating in the pairs with higher correlation value are highlighted.

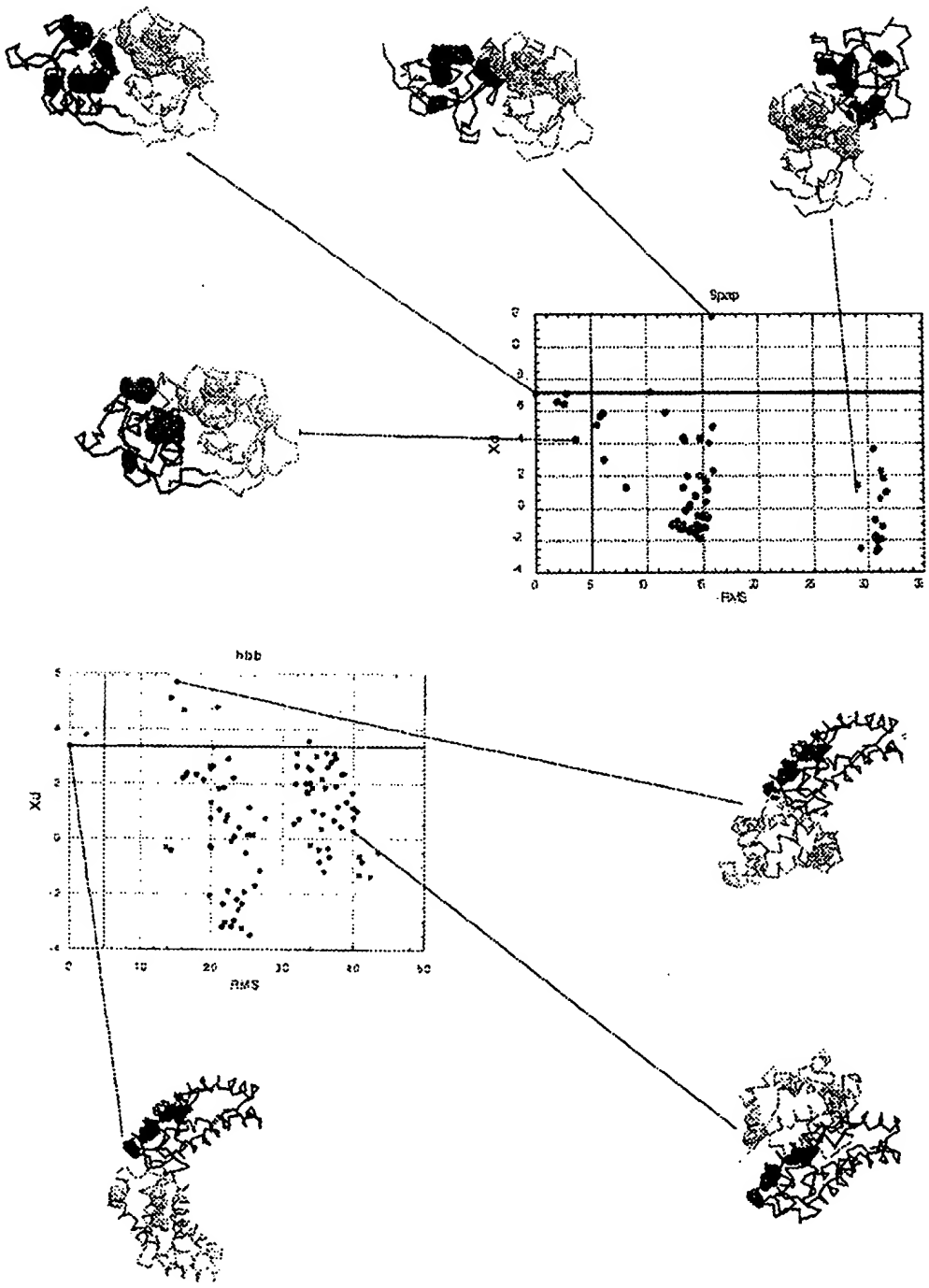

Figure 5. Scatter plot of the values of *Xd* against RMS for different docking solutions of (a) 9pap and (b) α1-β2 of haemoglobin (1hbb, Kavanaugh *et al.*, 1992). The *x*-axis, RMS between the real structure and the different alternative docking positions. The *y*-axis, harmonical difference (*Xd*) between the distributions of relative distances of all residues and correlated pairs. The right docking solution is among the 5% best scored ones. In the case of haemoglobin it is difficult to generate alternative docking solutions close to the real one, since the surface of interaction of the monomers is sharp and small. A ribbon representation of some docking solutions including the real one is shown. The residues participating in the pairs with higher correlation value are highlighted.

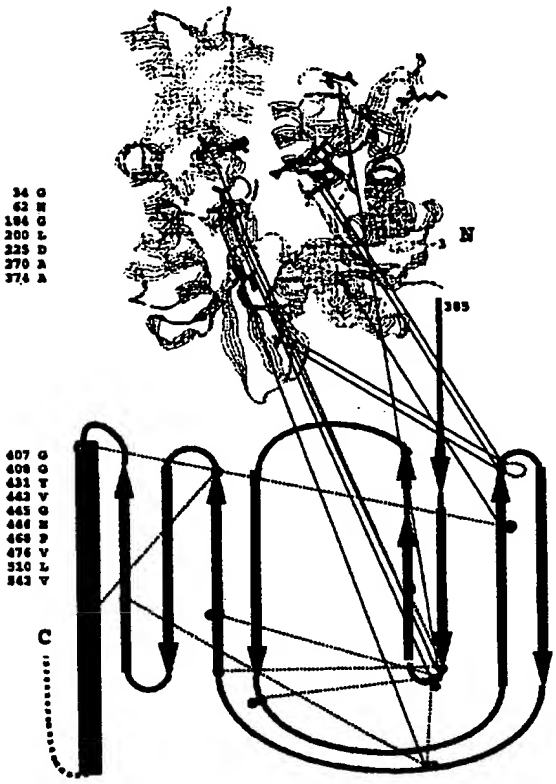

Figure 6. Predicted contacts between the Nt and Ct domains of Hsc70. In the upper part the three-dimensional structure of the Nt domain (3hsc, Flaherty *et al.*, 1990) is shown as a ribbon plot. A schematic view of the NMR secondary structure assignment of the Ct domain (Morshauser *et al.*, 1995) is shown in the lower panel. Strands are represented by arrows, helices by boxes. The residues undergoing correlated mutations between domains are shown as sticks in the ribbon plot. The ten best correlations are shaded and connected by lines. Their residue number and code are also given. Residues participating in the ten next best correlations are also shown as light sticks on the ribbon plot. Additionally, correlations between residues in the Ct domain are represented as broken lines. The Figure points to a clear docking solution between the two β sheets of the Ct domain and between the first β sheet of the Ct domain and the back of the Nt domains.

in the heat-shock protein Hsc70. The interaction between the Nt and Ct domains is essential for the function of the protein. The three-dimensional structure is known only for the Nt domain (Flaherty *et al.*, 1990: ribbon plot in Figure 6). The structure of th Ct domain has been solved recently for DnaK, a related protein (Zhu *et al.*, 1996), but it is not yet publicly available. A cartoon depicting the secondary and super-secondary structure of this protein as assigned by Morshauser *et al.* (1995) is shown in Figure 6.

The correlated mutations that we identified for this case clearly predict that two defined regions should be part of the interacting surface. Both regions map in the front face of the Nt domain (with respect to the standard view of Figure 6). This information could be directly tested by mutagenesis experiments and will be ultimately validated by the experimental determination of the complex between the Nt and Ct domains.

Discussion

The co-evolution of a protein–protein complex in different organisms must leave visible traces at the sequence level. Part of this information can be captured as correlated positions in multiple sequence alignments. We have previously shown that there is indeed a trend for correlated pairs of residues to be closer in space than non-correlated pairs of residues in single-domain proteins (Göbel *et al.*, 1994).

Here we verify that this behaviour is characteristic for residues in single domains (intra-protein contacts) and for residues sitting in two different protein domains (inter-protein). The information contained in our definition of correlated mutations is able to discriminate between the real docking solution and many other realistic but wrong alternatives in a significant number of cases. We have tested the method for two-domain proteins, for which it was possible to obtain a collection of examples. We anticipate that the same results hold for interacting proteins. Indeed, in the case of haemoglobin, correlated positions between the α and β chains are sufficient to single out the right orientation of the two domains among many alternatives.

We evaluated the performances of the method on known cases, and then used it to carry out a real prediction for the inter-domain interaction of the heat-shock protein Hsc70. This example illustrates the potential of the method to generate specific predictions about contacting residues and regions even when the protein structure is unknown.

Limitations of the method

The ability of correlated positions to discriminate between correct and incorrect relative positions of two domains is clearly related to the degree of physical proximity between the domains. Our interpretation is that only the co-evolution of closely interacting residues leaves detectable signatures at the sequence level.

The case of calmodulin is instructive, since correlated positions properly describe the domain contacts in the closed form of the protein (2bbm) without being biased by the existence of an open form. Therefore, predictions should be carried out only for interacting proteins, as is the case for Hsc70.

It is important to note that other factors influence the quality of our predictions: the quality of the alignment, the distribution of sequences in the family and the family size. As a rule of thumb, predictions are reliable only in families with more than 15 sequences. These sequences have to be well distributed, with both distant and close homologues. Since correlation is based on the co-adaptation of proteins, the analysis requires the alignment of co-evolved proteins. Although these data are not yet available for many protein families, the current pace of the different sequencing projects suggests that this limitation will be overcome very soon.

A further limitation of the method is that it is unreliable when applied to homo-multimers because it is impossible to distinguish between signals coming from inter or intramonomer contacts. As with NMR studies on homo-multimers, this problem can in principle be solved (O'Donoghue *et al.*, 1996).

Future prospects

Methods that use the three-dimensional structures of the proteins to be docked (Cherfils *et al.*, 1991; Fischer *et al.*, 1995; Helmer-Citterich & Tramontano, 1994; Jackson & Sternberg, 1995; Jiang & Kim, 1991; Shoichet & Kuntz, 1991; Stoddard & Koshland, 1992; Walls & Sternberg, 1992; Katchalski-Katzir *et al.*, 1992) are probably more accurate in the structural detail than that proposed here. However, our method has the clear advantage that it can be applied in the absence of any structural information, as we have shown here for Hsc70, and the prediction of contacts between domains could be a useful guide for experimental approaches even when structural information is not available.

It remains a major challenge to develop methods for detecting molecular partners using only sequence information. Correlated mutations may be used in this context, scanning data bases of multiple-sequence alignments for cases of compatible signals, presumably found in interacting proteins. For example, should we have a data base where each protein is represented by the same number of homologous sequences all from the same species, then we could in principle inspect the data base with a similar multiple alignment of the protein of interest and single out those proteins where the higher number of positions show a similar pattern of variation, i.e. those that have a higher number of correlated positions with respect to the query sequence.

It remains to be seen whether the development of such a method is feasible, but its existence would be extremely valuable for the various genome analysis projects, where complete cellular systems are described only by the sequence of their components and any procedure able to predict their network of interactions could be of enormous help.

M thods

Selection of a test set of two domain proteins

We have taken the two-domain proteins described by different authors (Holm & Sander, 1994; Siddiqui & Barton, 1995; Sowdhamini & Blundell, 1994; Swindells, 1994). From the initial set of 80 protein families, we left out those with less than 15 sequences in the HSSP data base (Sander & Schneider, 1993). Also discarded were those with many positions with gaps (positions with more than 10% of gaps are not included in the calculation of correlated mutations). Homodimers were also excluded, since it is impossible to distinguish between intra and inter-protein contacts. Our final list has 21 examples of two-domain proteins (given in Table 1 by their PDB identifiers, Bernstein *et al.*, 1977). We deliberately avoided manipulating the input data: multiple sequence alignments and domain definitions were taken directly from public sources. In the case of the haemoglobin α and β chains we have treated them as if they were a single protein with two domains by appending the sequences of the β chains to their corresponding α chains. Those species for which only one of the chains (α or β) is known were not included in the alignment. The final grand alignment contains 151 sequences coming from 147 species.

Calculation of correlated mutations and definition of correlation thresholds

Correlated mutations were calculated as described (Göbel *et al.*, 1994). Each position in the alignment is coded by a distance matrix. This position-specific matrix contains the distances between all pairs of sequences at that position. Distances are defined by the scoring matrix of McLachlan (1971). The association between each pair of positions is calculated as the average of the correlation for each corresponding bin of the position-specific matrices. Positions with more than 10% gaps or completely conserved were not included in the calculation.

The exact formula used in our calculation of the correlation coefficient (r_{ij}) for each pair of positions i and j of a protein with N proteins in its alignment is:

$$r_{ij} = \frac{1}{N^2} \sum_{kl} \frac{W_{kl}(S_{ikl} - \langle S_i \rangle)(S_{jkl} - \langle S_j \rangle)}{\sigma_i \sigma_j}$$

For each position in the alignment we have an $N \times N$ matrix where each element (k and l running from 1 to N) is the similarity (S_{ikl}) between the two residues (k and l) in this position (i) according to the given homology matrix. $\langle S_i \rangle$ is the mean of S_{ikl}, σ_i is the standard deviation of S_{ikl}.

Given that the accuracy of the predictions of contacts directly depends on the correlation values (Göbel *et al.*, 1994), the pairs of positions are sorted by their correlati n value and the top M residues

are defined as predicted contacts, with M proportional to the protein size. For this study, the number is set to half of the sequence length L, a compromise between accuracy of the prediction and the possibility of using a statistically significant number of correlated pairs. In practice the $L/2$ most correlated pairs of residues are split in three classes, domain I–domain I, domain II–domain II and domain I–domain II. The values given in Table 1 refer to these classes. In two cases no values are given in the Table because there were not enough pairs of residues among the $L/2$ best correlations.

Distance calculation and definition of the harmonic average (*Xd*)

We have previously used ACCURACY (number of correctly predicted contacts over total number of predicted contacts) to assess the reliability of predictions of contacts. ACCURACY is not the best measure in the case of domain–domain proximity, since we are looking for relative proximity between residues rather than for direct physical contact and in this case it is more reasonable to use a continuous measure of proximity. Distances between pairs of residues are grouped in bins of 4 Å and the distribution represented as relative proportions of pairs of contacts. Two different distributions of binned data are obtained for correlated pairs and for all pairs of positions. The difference between the two distributions is calculated bin by bin and weighted by a factor inversely proportional to the normalised distance (in Å) of the corresponding bin to increase the weight of closer distances.

Distances between residues correspond to C^{β}-C^{β} distances, C^{α} for glycine:

$$Xd = \sum_{i=1}^{i=n} \frac{P_{ic} - P_{ia}}{d_i \cdot n}$$

where, n is the number of distance bins (there are 15 equally distributed bins from 4 to 60 Å); d_i is the upper limit for each bin, e.g. 8 for the 4 to 8 bin (normalised to 60). P_{ic} is the percentage of correlated pairs with distance between d_i and d_{i-1}. P_{ia} is the same percentage for all pairs of positions. Defined in this way $Xd = 0$ indicates no separation between the two distance populations, $Xd > 0$ indicates positive cases where the population of correlated pairs is shifted to smaller distances with respect to the population of all pairs.

Generation of alternative docking solutions

To test if correlated positions contain information about protein-protein docking we compared the distance between correlated pairs of residues in the real structure with the distance in alternative docking solutions.

In the first experiment the full docking space was searched and a set of 7440 docking solutions were generated by rotating the second domain with respect to the first in 30° steps; for each orientation, ten random translations were generated to bring the two domains into contact (744 non-redundant domain I–domain II relative orientations and ten random translations for each one of them). For the fine-grain search around the real docking solution, a large number of alternative docking solutions were generated with a docking program called ESCHER (Ausiello *et al.*, 1997). Each protein is cut in 1.5 Å thick slices and the accessible surface of each slice is described as a polygon with 1.5 Å sides. The polygons representing the first protein are orderly superimposed to the polygons representing the second protein and the complementarity between them is evaluated. The evaluation of the geometric fit between the two surfaces depends on the number of sides that can be superimposed maintaining the corresponding vertices at a distance lower than a fixed threshold.

Complementarity is translated into a scoring scheme. A complete search in the rotation space is exerted by rotating the smaller protein in all possible orientations around to the bigger one. The cylindrical symmetry inherent to this kind of approach is very convenient in order to transform a three-dimensional surface matching problem into a simplified two-dimensional polygon comparison, but offers a very poor description of the target domain poles. In the solutions analysed here the target has been described only once with the interaction site parallel with the axis crossing the domain poles. In two cases (c2c and est) different sets of solutions were generated rotating the target protein 90° around the vertical axis. The efficiency of our method was similar considering one or more sets of solutions (not shown). For the purpose of this study the distance between the correct solution and alternative docking solutions is evaluated as the RMS deviation of the position of the second protein after superimposing the first one.

Acknowledgements

We are indebted to G. Cesareni, G. Casari, C. Ouzounis, U. Göbel and B. Rost for critical reading of the first manuscript draft. We also appreciate interesting discussions with Chris Sander. The help of Anna Tramontano and Sean O'Donoghue in the preparation of the final version and their scientific suggestions have been invaluable to us. The work of the Protein Design group CNB-CSIC in this area is financed by CICYT project BIO94-1067. ESCHER development has been supported by the Supercomputing Resource for Molecular Biology, Human Capital and Mobility Programme, Access to Large Scale Facilities grant, contract ERBCHGECT940062 and Telethon contract number 902.

References

Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. (1987). Correlation of co-ordinated amino acid sub-

stitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* **193**, 693–707.

Altschuh, D., Vernet, T., Berti, P., Moras, D. & Nagai, K. (1988). Coordinated amino acid changes in homologous protein families. *Protein Eng.* **2**, 193–199.

Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng.* **2**, 101–113.

Ausiello, G., Cesareni, G. & Helmer-Citterich, M. (1997). ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure. *Proteins: Struct. Funct. Genet*, In the press.

Babu, Y. S., Bugg, C. E. & Cook, W. J. (1988). Structure of calmodulin refined at 2.2 Å resolution. *J. Mol. Biol.* **204**, 191–204.

Bennett, M. J., Schlunegger, M. P. & Eisenberg, D. (1995). 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci.* **4**, 2455–2468.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Chappel, T. G., Konforti, B. B., Schmid, S. L. & Rothman, J. E. (1987). The ATPase core of a clathrin uncoating protein. *J. Biol. Chem.* **262**, 746–751.

Cherfils, J., Duquerroy, S. & Janin, J. (1991). Protein-protein recognition analyzed by docking simulation. *Proteins: Struct. Funct. Genet.* **11**, 271–280.

Fischer, D., Lin, S. L., Wolfson, H. L. & Nussinov, R. (1995). A geometry-based suite of molecular docking processes. *J. Mol. Biol.* **248**, 459–477.

Flaherty, K. M., DeLuca-Flaherty, C. & McKay, D. B. (1990). Three-dimensional structure of the ATPase fragment of a 70 K heat shock cognate protein. *Nature*, **346**, 623–628.

Fletterick, R. F. & Bazan, J. F. (1995). When one and one are not two. *Nature Struct. Biol.* **2**, 721–723.

Göbel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **18**, 309–317.

Gragerov, A., Zeng, L., Zhao, X., Burkholder, W. & Gottesman, M. E. (1994). Specificity of DnaK-peptide binding. *J. Mol. Biol.* **235**, 848–854.

Gregoret, L. M. & Sauer, R. T. (1993). Additivity of mutant effects assessed by binomial mutagenesis. *Proc. Natl Acad. Sci. USA*, **90**, 4246–4250.

Helmer-Citterich, M. & Tramontano, A. (1994). PUZZLE: a new method for automated protein docking based on surface shape complementarity. *J. Mol. Biol.* **235**, 1021–1031.

Holm, L. & Sander, C. (1994). Parser for protein folding units. *Proteins: Struct. Funct. Genet.* **19**, 256–268.

Ikura, M., Clore, G. M., Gronenborn, A. M., Zhu, G. & Klee, C. B. (1992). Solution structure of a calmodulin-target peptide complex by multidimensional NMR. *Science*, **256**, 632–638.

Jackson, R. M. & Sternberg, M. J. E. (1995). A continuum model for protein-protein interactions: application to the docking problem. *J. Mol. Biol.* **250**, 258–275.

Janin, J. & Chothia, C. (1990). The structure of protein-protein recognition sites. *J. Biol. Chem.* **265**, 16027–16030.

Janin, J., Miller, S. & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* **204**, 155–164.

Jayaraman, V., Rodgers, K. R., Mukerji, I. & Spiro, T. G. (1995). Hemoglobin allostery: resonance raman spectroscopy of kinetic intermediates. *Science*, **269**, 1843–1848.

Jiang, F. & Kim, S. H. (1991). "Soft docking": matching of molecular surface cubes. *J. Mol. Biol.* **219**, 79–102.

Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

Kavanaugh, J. S., Rogers, P. H., Case, D. A. & Arnone, A. (1992). High-resolution X-Ray study of deoxyhemoglobin Rothschild 37 beta TRP → ARG : a mutation that creates an intersubunit chloride-binding site. *Biochemistry*, **31**, 4111–4121.

Kamphuis, I. G., Kalk, K. H., Swarte, M. B. A. & Drenth, J. (1984). Structure of papain refined at 1.65 Å resolution. *J. Mol. Biol.* **179**, 233–256.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesen, A., Aflalo, C. & Vakser, I. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl Acad. Sci. USA*, **89**, 2195–2199.

Lengauer, T. & Rarey, M. (1996). Methods for predicting molecular complexes involving proteins. *Curr. Opin. Struct. Biol.* **5**, 402–406.

Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225–270.

McCarty, J. S., Buchberger, A., Reinstein, J. & Bukau, B. (1995). The role of ATP in the functional cycle of the DnaK chaperone system. *J. Mol. Biol.* **249**, 126–137.

McLachlan, A. D. (1971). Test for comparing related amino acid sequences. *J. Mol. Biol.* **61**, 409–424.

Montgomery, D., Jordan, R., McMacken, R. & Freire, E. (1993). Thermodinamic and structural analysis of the folding/unfolding transitions of the *Escherichia coli* molecular chaperone DnaK. *J. Mol. Biol.* **232**, 680–692.

Morshauser, R. C., Wang, H., Flynn, G. C. & Zuiderweg, E. R. P. (1995). The peptide-binding domain of the chaperone protein Hsc70 has an unusual secondary structure topology. *Biochemistry*, **34**, 6261–6266.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proc. Natl Acad. Sci. USA*, **91**, 98–102.

O'Donoghue, S., King, G. & Nilges, M. (1996). Calculation of symmetric multimer structures from NMR data using a priori knowledge of the monomer structure, co-monomer restraints, and interface mapping: the case of leucine zippers. *J. Biomol. NMR*, **8**, 193–206.

Pazos, F., Olmea, O. & Valencia, A. (1997). A graphical interface for correlated mutations and other structure prediction methods. *CABIOS*, **13**, 319–321.

Perutz, M. F. (1978). Hemoglobin structure and respiratory transport. *Sci. Am.* **239**(6), 92–125.

Rao, S. T., Wu, S., Satyshur, K. A., Ling, K. Y., Kung, C. & Sundaralingam, M. (1993). Structure of *Paramecium tetraurelia* calmodulin at 1.8 Å resolution. *Protein Sci.* **2**, 436–447.

Sander, C. & Schneider, R. (1993). The HSSP data base of protein structure-sequence alignments. *Nucl. Acids Res.* **21**, 3105–3109.

Serrano, L., Horovitz, A., Avron, B., Bycroft, M. & Fersht, A. R. (1990). Estimating the contribution of engineered surface electrostatic interactions to pro-

tein stability using double mutant cycles. *Biochemistry*, **29**, 9343–9352.
Shindyalov, I. N., Kolchanov, N. A. & Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations. *Protein Eng.* **7**, 349–358.
Shoichet, B. K. & Kuntz, J. I. D. (1991). Protein docking and complementarity. *Mol. Biol.* **221**, 327–346.
Siddiqui, A. & Barton, J. (1995). Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.* **4**, 872–884.
Sowdhamini, R. & Blundell, T. (1994). An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci.* **4**, 506–520.
Stoddard, B. L. & Koshland, D. E. (1992). Prediction of the structure of a receptor-protein complex usong a binary docking method. *Nature*, **358**, 774–776.
Strynadka, N. C., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B. K., Kuntz, I. D., Abagyan, R., Totrov, M., Janin, J., *et al.* (1996). Molecular docking programs succesfully predict the binding of a β-lactamase inhibitor protein to TEM-1 β-lactamase. *Nature Struct. Biol.* **3**, 233–239.
Swindells, M. B. (1994). A procedure for detecting structural domains in proteins. *Protein Sci.* **4**, 103–112.
Taylor, W. R. & Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**, 341–348.
Totrov, M. M. & Abagyan, R. A. (1994). Detailed ab initio prediction of lysozyme-antibody complex with 1.6 Å accuracy. *Nature Struct. Biol.* **1**, 259–263.
Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1996). Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. *Crit. Rev. Biochem. Mol. Biol.* **31**, 127–152.
Vernet, T., Tessier, D. C. & Khouri, H. E. (1992). Correlation of co-ordinated amino acid changes at the two-domain interface of cysteine proteases with protein stability. *J. Mol. Biol.* **224**, 501–509.
Walls, P. H. & Sternberg, M. J. (1992). New algorithm to model protein-protein recognition based on surface complementarity. Applications to antibody-antigen docking. *J. Mol. Biol.* **228**, 277–297.
Young, L., Jernigan, R. L. & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* **3**, 717–729.
Zhu, X., Zhao, X., Burkholder, W. F., Gragerov, A., Ogata, C. M., Gottesman, M. E. & Hendrickson, W. A. (1996). Structural analysis of sustrate binding by the molecular chaperone DnaK. *Science*, **272**, 1606–1614.

Edited by A. R. Fersht

(Received 1 April 1997; received in revised form 6 June 1997; accepted 6 June 1997)